

SYSTEM AND METHOD FOR ADJUSTING THE TRAFFIC CARRIED BY A NETWORK

Background of the Invention

The quality of service requirements for different types of information carried by a network are often themselves different. For example, it can be more important to limit the latency (the delay between the time a signal is sent and the time it is
5 received) and the error rate for voice information (e.g., conversations) carried by a network than for a file transfer carried by the same network. Even short delays in voice conversations lead to confusion because they can cause one party to mistakenly believe that the other party is finished speaking. This can cause one party to talk over the words of another party. Short delays in file transfers are typically
10 not noticed. Likewise, a high error rate in voice conversations can render a party's words incomprehensible. A high error rate in file transfers is typically handled without too much inconvenience using known error detection and correction protocols, and retransmissions.

A network typically has a finite capacity to carry information at a given
15 quality of service. Thus, meeting stringent quality of service requirements for one type of information can mean limiting the amount of other information carried by the network, or reducing the quality of service with which the other information is carried. One known scheme for carrying this out is the H.323 standard. The H.323 standard specifies gateways interposed between end-user terminals on a network.
20 The gateways act as "traffic cops," monitoring the state of the network and quality of service requirements, and deciding what information an end-user terminal can send

and receive, and by what parameters (e.g., rate, protocol, etc.) An H.323 gateway negotiates with another H.323 gateway to determine the parameters for a communication between end-user terminals represented by each gateway. This communication is typically performed on a link-by-link basis (e.g., it is performed
5 for each new communication between two or more end-users), and disadvantageously generates network traffic that can additionally burden a network. Further, a H.323 gateway's end-user terminals have no real control over the parameters of a communication, as these are dictated by the gateway.

Another known system involves direct negotiations between endpoints to
10 determine the parameters of a communication. Again, this is performed on a link-by-link basis, and can generate substantial amounts of network traffic when performed on a large scale (between many endpoints at once).

Yet another known system employs a gateway that monitors the status of the network and then acts as a filter between the gateway's end-user terminals and the
15 network, deciding what to pass on to the network and what to defer or drop, based upon the network's status (e.g., based upon the level of congestion of the network). The use of a gateway is expensive and adds an additional layer of complexity to network topology.

A better system would assure quality of service for a particular type of
20 information carried by a network without generating the substantial overhead traffic involved in two-way negotiations over communications parameters on a link-by-link basis, and avoid the expense and added complexity incurred by using gateways between the network and the end-user terminals.

Summary of the Invention

25 An embodiment of the present invention provides a system and method for improving or maintaining the quality of service delivered by a network with respect to a first type (e.g., a preferred type) of information carried by the network. The status of the network is monitored to determine if the actual or projected quality of service being provided by the network for the first type of information meets a

quality of service requirement. If the quality of service does not meet the requirement, a message is formulated and sent to an end-user terminal, such as a wireless handset. The message instructs the terminal to change the character of the information the terminal sends to the network. For example, the message can
5 instruct the terminal to send less of a second type (e.g., a non-preferred type) of information to the network, or change the compression scheme used to send any of the various types of information to the network.

Brief Description of the Drawings

FIG 1 shows an embodiment of the system and method in accordance with the
10 present invention.

FIG 2 is a flow chart that illustrates the method in accordance with the present invention.

Detailed Description

In accordance with an embodiment of the present invention, the quality of
15 service provided by a network with respect to a first type (e.g., a preferred type) of information is assured by directly regulating the character of information sent by end-user terminals to the network. As used herein, the "character of information" means any aspect of information that is sent from the end-user terminal to the network. For example, the character of information sent to the network from an end-
20 user terminal (such as a wireless handset) includes the type of information sent (voice information, video information, file transfer information, network signaling information, etc.); the relative quantity of one type of information in relation to at least one other type (e.g., 50% voice traffic and 50% electronic mail traffic, 35% video traffic, 60% voice-only traffic and 5% file transfer traffic), the compression
25 type(s) of information sent to the network, protocols used for the information, etc. As used herein, "changing the way an end-user terminal sends information to the network" is equivalent to "changing the character of information sent by the end-user terminal to the network."

In one embodiment of the present invention, the network is a wireless network (e.g., operating using the GSM or CDMA protocol), and the end-user terminals are wireless handsets. Examples of a wireless handset include a wireless telephone, a personal information manager (e.g., the Palm Pilot manufactured by the 3Com company of California) that is coupled to the wireless network, and any other appliance coupled to the wireless network. As used herein, the term "coupled" means directly or indirectly connected. Thus, A is coupled to B if A is directly connected to B. Likewise, A is coupled to B if A is directly connected to C and C is directly connected to B. In another embodiment, the end-user terminal can be a personal computer coupled to the wireless network.

An embodiment of a system in accordance with the present invention is shown in FIG 1. A wireless network 101 couples a network status manager 102 with end-user terminals 103 and 104. Other end-user terminals 105 and 106 are coupled to the network 101 through a gateway 107. In accordance with an embodiment of the present invention, the network status manager 102 monitors the performance of the network with respect to at least one type of information. The performance measurements by the network status manager 102 can be actual (i.e., the present performance of the network) or prospective (i.e., the projected performance of the network, which can be estimated using stochastic processes known in the art). For example, in one embodiment, the network status manager 102 measures the performance of the network 101 with respect to voice information by measuring the latency (the delay between the sending and receiving) of test voice packets sent by the network status manager 102 to various end-user terminals (e.g., 103, 104) or one or more gateways (e.g., 107) coupled to the network 101. This can be accomplished in one way by utilizing the Unix TRACE command well known in the art, which reports the times between hops by a packet across nodes in a network. Likewise, reporting mechanisms can be set up between end-user terminals (e.g., 103 and 104) and the network status manager. The reporting mechanisms can generate periodic reports pertaining to network performance for one or more types of information

based upon observations made at the end-user terminals (103, 104, etc.) These observations can include latency measurements, packet loss rates, etc.

The network status manager 102 monitors the performance of a wireless network with respect to voice (e.g., conversation) information in one embodiment, although the wireless network also carries non-voice information, including e-mail messages, world wide web information (information exchanged in the course of browsing web sites on the world wide web, such as http requests and web site data), etc. In this embodiment, the voice information (the first type of information) is identified as "preferred information" type by the network status manager 102, and the other types (second types) of information carried by the network are designated "non-preferred" types of information. In this case, this is because delays and errors can have a more disruptive impact on the quality of voice communications than on the exchange of e-mail or world wide web information. For example, when the network becomes congested and a download of e-mail is delayed for a few seconds, a user is typically not substantially inconvenienced. But if a party's voice information is delayed for a few seconds, then the conversation can be substantially disrupted. For example, one isn't sure when the other party has stopped talking, which can lead to talking over the other person. Also, if there is a high error rate in carrying e-mail traffic, the e-mail can simply and quickly be resent until it is received correctly. On the other hand, a high error rate for voice communications can render words or sentences incomprehensible.

In accordance with an embodiment of the present invention, the network status manager identifies a minimum quality of service requirement for the preferred type of information. For example, the network status manager maintains a maximum latency (delay between the time a piece of data is sent and the time it is received) and a maximum error rate for voice information. If the (current or projected) performance of the network with respect to the preferred type of information does not meet the quality of service requirements for that type of information, then the network status manager formulates a network status message that it sends to at least one of the end-user terminals. In accordance with an embodiment of the present

invention, the network status message is adapted to instruct the end-user terminal to change the way it sends information to the network in order to alleviate whatever problem is causing the network to fall short of the quality of service requirement for the preferred type of information. Thus, for example, if the quality of service
5 requirement for voice traffic in a network is a maximum average latency of no more than 500 milliseconds, and the average latency for voice traffic is determined to be 750 milliseconds, then the network status manager sends a network status message to at least one end-user terminal to correct this situation, i.e., to lower the average voice packet latency to below 500 milliseconds. This can be accomplished in several
10 different ways.

In one embodiment, the network status message instructs the end-user terminal to change the compression scheme it uses for the preferred type of information. For example, the network may be receiving too much voice information to maintain an adequate voice quality of service, so the network status manager
15 formulates a message to end-user terminals to use a more efficient compression scheme for sending voice information to the network. Thus, whereas a first compression scheme sends a second of speech using 5000 bytes of information (a byte is 8 bits, e.g., 10110011), a second compression scheme sends a second of speech using only 2000 bytes. Thus, when the network status manager 102
20 determines there is too much congestion on the network, the network status manager sends a network status message to handsets coupled to the network to switch from the first compression scheme for voice to the second compression scheme. The more a voice signal is compressed, the worse it typically sounds, so using the more efficient second compression scheme should typically only be mandated when
25 necessary to preserve overall quality of service. A lower quality voice signal may be an acceptable tradeoff in order to prevent excessive delays in receiving fragments of conversation.

In another embodiment of the present invention, the compression scheme of non-preferred information is altered by the receipt of a network status message by an
30 end-user terminal. For example, an end-user terminal is required to use more

efficient compression in sending video images to the network. This can advantageously free up bandwidth to maintain a high quality of service for the preferred information type.

In another embodiment of the present invention, the rate at which
5 information is sent to the network is changed by the receipt of the network status message. In one embodiment, the network status message directs an end-user terminal to slow the rate at which it sends non-preferred information (e.g., e-mail information) to the network, so as to free up bandwidth for maintaining the quality of service of preferred information (e.g., voice information). In another embodiment,
10 the network status message directs an end-user terminal to stop sending any new non-preferred information to the network until further notice. In yet another embodiment, the network status message specifies a maximum rate at which non-preferred (and/or preferred) information can be sent to the network. For example, the network status message can instruct the end-user terminal to send file transfer
15 information at a rate no greater than 200 bytes per second.

The network status message is sent as a signal to an end-user terminal, typically (although not necessarily) through the same network whose quality of service is being monitored by the network status monitor 102. In another embodiment, the network status message is sent to the end-user terminal through a
20 second network (not shown). In one embodiment, the network status message signal includes the following:

- First Bit: enable/disable transmission of voice information
- Second Bit: enable/disable transmission of e-mail information
- Third Bit: enable/disable transmission of video information
- 25 Fourth Bit: enable/disable transmission of http requests
- Fifth Bit: enable/disable transmission of ftp requests

For example, the signal 11010 enables an end-user terminal to send voice information, e-mail information and http requests to the network, and prohibits the end-user terminal from sending video information and ftp requests to the network.

Another example of a network status signal includes the following:

- 5 First Two Bits: specifies compression scheme for voice information
- Second Two Bits: specifies compression scheme for file transfer
- Third Two Bits: specifies compression scheme for video information.

Thus, the signal 101100 in accordance with this example specifies compression format 2 for voice (the binary digits 10 are equivalent to the number 2 in the decimal system); compression format 3 for file transfer, and compression format 0 for video information. Each of the compression formats for each of the types of information are advantageously predetermined and known by the handset. Thus, for example, the handset knows that compression format 0 for file transfer is the Kermit data compression format, compression format 1 is the ZIP compression format, etc. Note
10 that one of the compression formats specified in accordance with an embodiment of the present invention is no compression. For example, the compression format 0 for video information in the above example can specify that no compression is to be used for video information sent to the network. This can be advantageous when the clarity of video images is especially important.

20 Yet another example of a network status message signal includes:

- First Four Bits: specifies the rate at which voice information can be sent to the network
- Second Four Bits: specifies the rate at which e-mail information can be sent to the network
- 25 Third Four Bits: specifies the rate at which video information can be sent to the network

For example, the signal 101110010100 indicates that voice information can be sent at predetermined rate number 11 (11 is the decimal representation of the binary number 1011); e-mail can be sent at the predetermined rate number 9; and video information can be sent at predetermined rate number 4. These rates can be

5 measured differently for each type of information. For example, rate 11 for voice can mean 5000 bytes per second, rate 9 for e-mail can mean 9600 bits per second, and rate 4 for video can mean 8 frames of video information per second. The predetermined rates may also represent maximum rates, thereby permitting the end-user terminal to send the information at a slower rate.

10 The above examples of network status message signals can be advantageously combined to specify, e.g., the enablement, compression and rate at which one or more types of information can be sent from the end-user terminal to the network.

Another advantageous feature of the present invention is the ability to
15 selectively transmit network status messages only to certain end-user terminals, and to transmit different network status messages to different terminals. This is especially useful when the network is an internetwork (a set of networks coupled to each other), or when the source of network congestion is localized. For example, it may be disadvantageous to limit the transmission of all e-mail to the network from
20 all end-user terminals, or to specify a more efficient voice compression scheme for all end-user terminals when the source of network congestion is originating from a specific subset of terminals. The preferential formulation and/or transmission of network status messages can advantageously be performed by any criterion aimed at solving the problem that jeopardizes the network's capability to meet a quality of
25 service requirement, while at the same time minimizing the negative impact of the network status message (e.g., lowering voice quality by mandating a more efficient voice compression scheme). For example, the network status messages can be formulated and/or sent on the basis of geography (e.g., to all end-user terminal in Minneapolis, to all terminals connected to a given cell in a cellular wireless network,
30 etc.), on network topology (to all terminals connected to a given subnetwork, to all

users that utilize a given protocol, to all users with a given area code, etc.), and so on.

For example, a sporting event attended by a large number of persons with cell phones routinely causes a large amount of voice data to be sent to the network from terminals in the area of the stadium in which the event is held. Without using the network status message in accordance with an embodiment of the present invention, the local part of the network would become congested, resulting in poor voice quality and/or denial of service to many customers. In accordance with an embodiment of the present invention, a network status message is sent only to terminals connected to the part of the network in the stadium (e.g., to the handsets communicating with the cell that serves the stadium) to mandate the most efficient compression scheme for voice, and prohibiting the transmission of any other type of information to the network. The congestion problem is thereby solved without impacting other users whose terminals are not coupled to the stadium part of the network. In this way, the present invention is capable of adjusting the way (compression, rate, etc.) in which information is sent to the network with a minimal adverse impact on the smallest number of customers.

An embodiment of an apparatus (a network status manager) in accordance with the present invention is illustrated in FIG 1. A processor 201 is coupled to memory 202, which stored network status instructions 203 adapted to be executed by processor 201. Processor 201 in one embodiment is a general purpose microprocessor, such as the Pentium III microprocessor manufactured by the Intel Corporation of Santa Clara, California. In another embodiment, processor 201 is an Application Specific Integrated Circuit, which is designed to implement in hardware and firmware at least a part of the method in accordance with an embodiment of the present invention. Memory 202 is any device capable of storing digital information, such as a Random Access Memory, Read Only Memory, flash memory, a hard disk, a read/writable magneto-optical disk, etc. The apparatus shown in FIG 1 also includes a port 204 adapted to be coupled to a network 101. Port 204 is coupled to processor 201. Network status manager instructions 203 are adapted to be executed

by processor 201 to perform the method in accordance with an embodiment of the present invention, such as the one shown in the flow chart depicted in FIG 2.

FIG 2 is a flow chart showing the method in accordance with an embodiment of the present invention. A quality of service requirement for a type of information carried by the network is determined, 301. Network performance is determined, 302. It is determined if the network performance meets or exceeds the quality of service requirement, 303. If so, then no action is taken, 304. If it is determined that network performance does not meet the quality of service requirement, 303, then a network status message is formulated, 305. The network status message is adapted to be received by an end-user terminal coupled to the network to change the way the terminal sends at least one type of information to the network. The network status message is then sent to an end-user terminal, 305.

The present invention advantageously manages network quality of service by controlling the manner in which end-user terminals are permitted to send information to the network. The embodiments described above are meant to illustrate, and not limit, the scope of the present invention. For example, the first and second types of information need not correspond to "preferred" and "non-preferred" types. The first and second types can bear any relation to each other. Further, the adjustments made as a result of the comparison between the quality of service requirement and performance of the network with respect to a type of information encompass any change in the character (e.g., amount, type, timing of, compression scheme of, encoding of, etc.) any of the various types of information sent to the network. An embodiment of the present invention operates by sending instructions to end-user terminals, such as wireless handsets, computer terminals communicating using wireless telecommunications technology, handheld devices such as personal data managers, etc., to change the way they send information to a network based upon the quality of service requirement and network performance with respect to at least one type of information carried by the network. These and other aspects are encompassed by the present invention, as will be apparent to those skilled in the art.